DOCUMENT RESUME

ED 064 364 TM 001 532

AUTHOR Hess, Robert J.; Wright, William J.
TITLE Evaluation Strategies as a Function of Product Development Stages.
INSTITUTION Central Midwestern Regional Educational Lab., St. Ann, Mo.
SPONS AGENCY Office of Education (DHEW), Washington, D.C.
PUB DATE Apr 72
NOTE 30p.; Paper presented at the American Educational Research Association Convention (April 1972, Chicago, Ill.)

EDRS PRICE MF-$0.65 HC-$3.29
DESCRIPTORS Consumer Education; *Curriculum Development; Economic Factors; *Educational Research; *Evaluation Techniques; Federal Government; Measurement Instruments; Program Evaluation; *Standards; Testing

ABSTRACT

There are issues in evaluation and stages of product development that demand the use of experimental or quasi-experimental designs. To counteract criticism of evaluation efforts, an approach to the examination of the multiple issues involved in curriculum product evaluation across the usual developmental cycle of educational products was developed. Curriculum products typically move through a developmental sequence comprised of five stages: Initial State, Hot House - the initial tryout of a prototype product, pilot test, field test, and public diffusion. Each stage represents a milestone in the life of a product. In the course of evaluation, various audiences are acquired: the sponsor, the institution, the developer, consumer representatives, and advisors. There are five major dimensions of a comprehensive evaluation of curriculum products: Desirability/Feasibility, (2) Management/Procedural Cost, Product Worth, Usability, and Generalizability. Issues relating to the continuation or termination of a program concern statement and fulfillment of objectives, establishing a rationale for the use of particular measuring instruments, determination of whether or not different effects result from alternative procedures. When the product enters the diffusion stage, formative evaluation is ended and summative evaluation ought to begin. It is pointed out that true summative evaluation is consumer protection and is a three tiered process operating wherein: (1) The product developer establishes the criteria; (2) some agency of the federal government examines the product; and (3) Local education agencies research the products. (CK)

Evaluation Strategies as a Function of Product Development Stages

Robert J. Hess and William J. Wright

CEMREL, Inc.

Published by CEMREL, Inc., a private non-profit organization supported in part as a regional laboratory by funds from the United States Office of Education, Department of Health, Education, and Welfare. The opinions expressed in this publication do not necessarily reflect the position or policy of the Office of Education, and no official endorsement by the Office of Education should be inferred.

Paper presented at the American Educational Research Association Convention, April 1972, Chicago, Illinois

INTRODUCTION

The recent past has witnessed the emergence of curriculum evaluation as a critical concern among the educational research and other communities. The major impetus for this concern has come from various curriculum projects such as that of the Physical Science Study Committee and the funding of laboratories and centers for curriculum research and development by the U. S. Office of Education. Models for evaluation activities in connection with curriculum development activities have been proposed by scholars in various disciplines, and we now appear to be at that point in time where it becomes possible to attempt a synthesis of some of these ideas into a coherent system for the evaluation of curriculum products.

Evaluation, as the most salient term in this paper, requires definition. The recent literature (especially Guba and Stufflebeam [1968], Stufflebeam [1971], MacDonald [1970], Astin and Panos [1971]) has tended to foster a view of curriculum evaluation in terms of its role in the realm of educational decision-making. In this role, curriculum evaluation may be defined in the manner of Guba and Stufflebeam (1971) as "The process of obtaining and providing useful information for making educational decisions (pp. 23-24)." The critical terms are "useful" and "decision." We would argue (with Guba and Stufflebeam [1968], Astin and Panos [1971], and others) that unless the information gathered and reported will be used in the determination of which among alternative options will be chosen, then there is little or no reason to engage in evaluation.

The issue that has perhaps been most pervasive of late in the field of curriculum evaluation was raised most succinctly in Scriven's (1968) AERA monograph paper. In dealing with the roles and goals of evaluation, Scriven posited the distinction between formative and summative evaluation. Briefly, formative evaluation refers to those activities that provide information to the developer concerning his product's worth and effectiveness when the product is still fluid and most amenable to change; summative evaluation takes place when a product is a completed entity.

The recent literature has dealt extensively with the formative-summative dichotomy (e.g., Grobman, 1968; Cronbach, 1963; Karl, 1970; and Wittrock, 1966). As Weiss (1971) has noted, however:

> In contrast with the large body of literature concerned with measuring and diagnosing student progress is a lack of clear cut systematic procedures for the logical and judgmental questions of formative curriculum evaluation. (p.4)

Moreover, some individuals have viewed the distinction as being one of rigor rather than one of role. Scriven, himself, refuted that interpretation when he writes, "If you think that formative evaluation can be informal and that only summative needs to be done with any kind of rigor . . . then you find yourself simply making the wrong intermediate decisions (1971, p.1)."

Marion Karl (1970), in his explication of process evaluation, (". . . evaluation that illuminates areas where revisions are necessary while the opportunity for revision still exists"), notes that few attempts are made to conduct evaluation of this type and attempts to explain this lack.

> No neat, scientific research design is probable. Assumptions inherent in the models for statistical analysis of the data must be violated. The researcher doesn't have a tenth of the control he would like. Subjectivity is rampant. At times, one feels he is attempting a task as impossible as analyzing the water at a given point in a moving stream. (p. 1)

Indeed, the timely distinction between formative and summative evaluation has served to focus attention on the suitability of the application of experimental or quasi-experimental designs (Campbell and Stanley, 1963) as the sine qua non of evaluation methodology. During the course of the development of the product some of the purposes of the evaluation activies may not be achievable by the application of experimental methodology. As Guba and Stufflebeam (1968) have noted:

> On the surface, the application of experimental design to evaluation problems seems reasonable, since traditionally both experimental research and evaluation have been used to test hypotheses about the effects of treatments. However, there are four distinct flaws with this reasoning.
>
> 1. The application of experimental design to evaluation problems conflicts with the principal that evaluation should facilitate the continual improvement of a program . . .
>
> 2. Experimental design is useful for making decisions after a project has run full cycle, but almost useless as a device for making decisions during the planning and implementation of a project . . .
>
> 3. Experimental design is suited to the antiseptic conditions of the laboratory but not to the septic conditions of the classroom . . .
>
> 4. While internal validity may be gained through the control of extraneous variables, such an achievement is accomplished at the expense of external validity . . . (p. 14-16)

As we hope to demonstrate later in this paper, there are issues in evaluation and stages of product development that demand the use of experimental or quasi-experimental designs, but to require that all evaluation activities

conform to this single set of standards is to adopt a procrustean solution to an array of complex problems. There are legitimate questions, critical to the development of a curriculum, for which experimental designs are inapplicable. In short, we argue that formative evaluation needs to be formal and rigorous, but need not in all instances be equivalent to experimental research.

There has been some criticism of evaluation efforts based on the grounds that they examine only student learning outcomes and usually only outcomes that fall within a narrow band, i.e., those that can be behaviorally specified a priori. It strikes us that this criticism is just and warranted. It is no doubt true that the ultimate worth of most educational products rests largely on whether they are effective agents for the promotion of student learning, yet there are other important criteria, not only ultimate but intermediate and emergent. Considerations of intermediate and emergent issues necessitate a dynamic view of curriculum evaluation. As Grobman (1968) has stated:

> . . . the curriculum itself is emergent. As curriculum ideas work, they are pursued further; as they fail, they are modified or scrapped. In the same way, the evaluation is emergent; it must adapt to the readily changing experimental curriculum and to the changing needs of the project . . . perhaps emergent and dynamic are the two best adjectives to describe curriculum project evaluation. (p. 10)

Indeed, it may be argued that it is incumbent upon the evaluation program to provide information <u>beyond</u> whether or not particular learning-outcome goals were achieved, since such information is at best only minimally sufficient for decisions concerning adoption or rejection by a school system or other "purchaser." Again, borrowing from Grobman (1968) we may note:

> A curriculum assessment that concerns itself only with the instructional materials, without some understanding of the other variables in the situation, may conclude that certain results are produced, but may give no indication of why these are produced, or why different results ensue in different situations. Or it may conclude that the materials have failed, when the failure reflects a different factor. (p. 5)

Tyler and Klein (1970), in their recommendations[1] formulated for curriculum and instructional materials, have also recognized the importance of the nature and extent of information that curriculum developers should provide for consumers regarding their packages, since

> At one time curriculum and instructional materials were made locally by various publishers, but were used in a limited manner. If the materials were inadequate, the harm was restricted. This is not so likely today, and with the merging of electronic organizations and publishers, it is a certainty that curriculum and instructional materials will be centrally made and widely used. The damage could be widespread. (p. 1)

STAGES OF DEVELOPMENT/EVALUATION[2]

In order to attend to the variety of concerns raised by those cited above and others, what may be needed is an approach to the examination of the multiple issues involved in curriculum product evaluation across the usual developmental cycle of educational products. It has been our experience in working with curriculum development projects at a national educational laboratory (CEMREL) that curriculum development projects typically move through a developmental sequence comprised of five stages.

---

[1]These recommendations parallel those developed by the APA Committee on Test Development regarding the nature and extent of the information to be provided in manuals accompanying curriculum packages.

[2]The stages we delineate are obviously evaluation points in the life of a product. Developers undertake a great deal of development and revision between these events, and perceive that effort as almost the totality of the development effort. One way to state the case we are making is that development isn't a sprint, and the stages we pose are the hurdles on the track.

We will characterize these stages here as:

1. Initiation Stage - a general specification of what the curriculum project intends to do, how it is to be done, and for whom it is to be done.

2. Hot House - The initial tryout of a prototype product, typically in one or two classrooms with teachers who have a continuing relationship with the program.

3. Pilot Test - A systematic, small scale trial of the revised product, generally in proximate school systems. The teachers of these classrooms have access to staff and resources not expected to be available to the eventual users.

4. Field Test - Extended use of the ultimate or penultimate product in sites removed from the development institution. Program development staff serves no mediating role; the product is "on its own."

5. Public Diffusion - The product is commercially published in large quantity and is available to interested consumers.

These stages are, of course, arbitrary. In dealing with any real product, development must be a more flexible process. Certain material may require recycling through parts of the process, other material may need only a truncated portion of the sequence due to the limited goals held for the product. For our purpose though, it is useful to identify a delimited set of stages; at the same time, we acknowledge that our suggestions require adaptation in any real setting.

Each stage in the developmental/evaluative sequence we have outlined represents a relatively easily identified milestone in the life of a curriculum product. Moreover, the entry of a product into a given stage can be conceived as a decision point. If a product lacks certain characteristics or has failed to meet certain ends required for its next stage, one can terminate development or go into a recycling loop if that seems the most desirable course of action. (Obviously this statement over

simplifies the complex realities involved in any pragmatic decision-making model. Nevertheless, it is fair to say that we view the criterion-acquisition-before-advancement notion as the foundation on which the process should be built.)

THE AUDIENCES OF EVALUATION

In addition to identifying the stages of product development, it is necessary to characterize the various audiences for the information acquired by evaluative activities.[3] The audiences we have identified fall into five general categories.

The Sponsor -

Within this group are the public and private support organizations who provide the financial resources necessary for development; the U.S. Office of Education, state departments of education, foundations, and the like. The major interests of this group are perceived to be the identification of educational needs and the monitoring of development activities to insure that the products will have substantial impact in the natural setting of the schools within an established time limitation.

The Institution -

The funded organization which allocates resources to the development program. The central administration of regional laboratories and research and development centers and their boards of directors would fall into this classification. The major concern here is to insure that the

---

[3]Much of what follows has been adapted from "Evaluation of Educational Research and Development Activities," a document prepared by the CEMREL Evaluation staff and which served to promote the discussion of these issues at the 1970 meeting of CEMREL's National Advisory Committee on Evaluation.

management and direction of the project are fundamentally sound, and that the work undertaken is in accord with the established contracts and the mission of the institution. This group also has a vested interest in the quality control of products developed under their auspices.

The Developer -

May be equivalent to the institution if it is a single-mission organization. In many settings, however, there exists a staffing arrangement wherein a project developer is assigned responsibility for a given product or set of products comprising a portion of the total expected output of the organization. The developer is the person with the greatest variety and breadth of decision-making accountability. It is his duty to make the fundamental decisions concerning goals, the means of achieving them, the allocation of resources within the project, product sequencing, and quality control.

Consumer Representatives -

This group is made up of those who will eventually use the products or be immediately affected by their use - superintendents, principals, teachers, students, parents, and the public community to which the schools are responsible. The pre-eminent issues for this category are cost, pupil outcomes and teacher usability.

Advisors -

We have divided this category into two sub-groups in line with the functional arrangement at CEMREL. The substantive advisory group represents expertise in the content discipline of the programs. They serve to aid the developer in the specification of goals and means, and to periodically review the work in progress using intrinsic evaluative criteria.

The evaluation advisory group is comprised of prominent methodologists and researchers who review the procedures used to evaluate products and suggest alternative approaches. Their major interest is the assurance that adequate precautions have been taken to promote quality products through the collection of valid, reliable information about their use and effects.

Once again we note that we do not pretend to have encompassed all relevant audiences. Any given product development effort must specify its own audiences and their roles if it is to adopt our suggestions appropriately.

CRITERION ACQUISITION FOR PRODUCT ADVANCEMENT

In a recent article, Stufflebeam (1971) examined the utility of experimental designs for a variety of questions pertinent to evaluation studies using the four elements of the CIPP evaluation model (Context, Input, Process and Product) as question categories. Building on this notion, we have identified five major dimensions of a comprehensive evaluation of curriculum products. We have attempted to illustrate the types of questions that fall within each dimension. The tables that follow juxtapose these questions against the five development stages identified above. Within the cells of the tables we have identified criteria to be met before a product moves to the next development/ evaluation stage, as well as the audiences for whom evidence of criteria acquisition might be considered relevant. There are, of course, virtually unlimited modifications that could be made in these tables for a real product development effort.

INSERT TABLES I-V

The major dimensions we have identified for the evaluation of a curriculum development enterprise are, in general, the dimensions utilized by CEMREL to serve as a guide for the development of evaluation activities. These dimensions follow, along with a brief, general description of the particular foci of concern.

I. Desirability/Feasibility (Table 1) - Issues of interest pertinent to this dimension are those typically concerned with establishing a sound rationale for the justification of the commitment of the resources necessary to fulfill a consumer need of recognizable import.

II. Management/Procedural/Cost (Table 2) - Issues of interest here are those typically concerned with the administrative and resource allocation components of the product development enterprise.

III. Product Worth (Table 3) - This dimension deals with issues related to a specification of the nature of the product, it's accoutrements (i.e., materials, assessment devices, etc.) and the effects resulting from product use.

IV. Usability (Table 4) - This dimension is typically concerned with the use of the product by a sample of the target population so as to characterize and specify implementation strategies of known worth.

V. Generalizability (Table 5) - The foci of interest here are typically concerned with evaluating the evaluations and new applications of the product.

Inspection of these Tables (1-5) will reveal some of the relevant issues subsumed under each dimension that could conceivably be issues of critical concern for a given product development program. Notice that while many of the issues identified have specified criteria for each stage (i.e., milestone) in the developmental sequence, other issues may require some intermittent criteria sequence (see issue F, Table 3), or a sequence that becomes emergent at some later stage in the developmental sequence (see issue B, Table 5). Other issues may be served by a sequence that terminates prior to termination of the program (see issues A, B, and C, Table 1). The achievement of specified criteria serve as the determinants of milestone acquisition in the development of program components (issues) of interest. These milestones in the developmental sequence are major decision points for determining whether or not certain initial program components may continue through the sequence, must be revised and recycled to attain specified criteria, or perhaps terminated if the criteria cannot be met.

Decisions concerning continuation, revision and recycling, or termination of program components relevant to the milestone sequence are made by those members of the audience who decide whether or not specified criteria have been met. Audience composition is determined largely by the relevance of audience expertise for the evaluation of criteria acquisition as well as the relative impact these issues and criteria may have for the total program and/or the institution administering the program or those sponsoring the program. Consumer representatives are included in the audience in almost all instances where judgements of criteria acquisition for relevant program components immediately precede public diffusion.

To illustrate the relationships between the milestone stages in the developmental sequence and program components of interest, we shall choose a major dimension and discuss the relationships among issues, criteria, milestones, and audiences. The dimension we have chosen is that of product worth (Table 3) since this dimension might be of major concern to the present audience. We shall consider each issue separately and discuss the criteria and relevant audience as specified for each milestone in the developmental sequence.

The first of the issues identified under the product worth dimension is, "Are the objectives clearly stated?" The criterion to be met prior to the necessary commitment of resources (e.g., allocation of money for hiring substantive staff to develop program units, support services, overhead, materials, etc.) involved in program initiation would appear to demand (from the developer) some specification of program objectives in terms of some desired student acquisitions (e.g., skills, attitudes, knowledge, etc.). The most relevant audience attending to this criterion would presumably be comprised of the substantive advisors, since the objectives are directly relatable to substantive issues; the sponsor, since he is presumably concerned over what will result from his investment; and the institution, since it typically must share responsibility with the developer for program outcomes and would, therefore, demand some influence over statements concerning program objectives.

Prior to entering hot house (the initial input of a prototype product in one or two classrooms) the objectives should be stated in terms relatable to observable phenomenon to enable the initial development of measurement

instruments to determine whether the objectives have been met. The most relevant audience here would be comprised of the evaluation advisors, since they are specialists in measurement; and the substantive advisors, whose task here is to insure the viability and clarity of the stated relationship between the objectives and the phenomenon.

Since revisions of curriculum packages/units resulting from experiences in hot house are the rule rather than the exception, the criterion to be met prior to the pilot test (i.e., systematic, small scale trial of the revised product in proximate school systems) is to insure that any revisions of objectives are again stated in terms relatable to observable phenomenon, hopefully yielding more precision in statements concerning these relationships. The audience again is comprised of the evaluation and substantive advisors for the reasons stated previously. Presumably no further revisions are necessary beyond this point, however, if necessary, the same criterion could be applied prior to the field test.

The second issue we have identified asks, "Are the objectives being met?" Since the answer to this question typically requires the utilization of measurement instruments in the context of some assessment strategy, no criteria could seem to be applicable prior to hot house when measurement devices have been developed. Consequently, the criterion we have identified to be met prior to hot house is the approval by the Evaluation Advisory Committee of the general assessment strategy proposed by the evaluation staff serving the program. Since the over-all assessment strategy is a critical element of the program presumably involving commitment of substantial resources and being an issue of major concern, the audience we have identified as being most relevant to this criterion is comprised of the

evaluation advisors, the sponsor, the institution, and the developer, whose case rests on the results yielded from such an assessment strategy.

The criteria to be met prior to entering the pilot test have been identified as: 1) providing evidence that postulated outcomes were achieved in hot house, and 2) the establishment of an experimental design to test experimental hypotheses of interest regarding the objectives. Underlying these criteria is the assumption that relevant aspects of the overall assessment strategy had been implemented in hot house to yield evidence about the postulated outcomes. Presumably all audiences (i.e., evaluation advisors, substantive advisors, sponsor, institution, developer, and consumer representatives) would be interested in the evidence as presented and the nature of the information to be yielded about the objectives from the experimental design to be used in the pilot test.

The criteria to be met prior to the commitment of resources necessitated by a large scale field test require the verification of the experimental hypotheses from the pilot test and the establishment of some means to acquire relevant data in disparate field sites. Again, all audiences are seen as being relevant for evaluating the attainment of these criteria due to: 1) the resources necessary for field tests (sponsor), and 2) the fact that it is the first reasonably large scale test of the most critical aspect of the product (developer, substantive advisors, evaluation advisors, institution, and consumer representatives).

Prior to public diffusion, there should be evidence that the experimental effects noted in the pilot test and tested in the field test are generalizable across the field test sites. All audiences are again

perceived as being relevant due to the critical nature of claims regarding the attainment of objectives.

The next issue we have identified is concerned with establishing a rationale for the use of particular measuring instruments for purposes of evaluating the attainment of product objectives. Since this issue directly corresponds with the initial assessment of objectives during hot house, no criterion would seem to be applicable at the initiation stage of product development. Prior to the use of measures in hot house, the criterion to be met is that of establishing the content validity of the measures. (Evidence regarding other aspects of validity and issues of reliability may not be resolvable at this point due to the limited population samples involved.) The most relevant audience for this criterion would be comprised of the evaluation advisors (for their measurement expertise) and the developer (again since his case rests primarily on the measures used to evaluate his product).

The criterion to be met prior to the use of any measures in the pilot test consists of the accomplishment of necessary revisions (as indicated during hot house observations) of the data collection instruments (e.g., item additions or deletions, variations in item format, etc.) or techniques (e.g., weekly as opposed to bi-weekly or daily tests, etc.). The relevant audience here would be comprised of the evaluation advisors and the developer with the addition of the substantive advisors (since the measures are typically substantially bound and are in the process of being finalized) and the institution (since data collection and analysis procedures may require support services and involve subject protection considerations).

Since an adequate pilot test should involve a fairly large number of subjects, data concerning the validity and reliability of the measures used would be derivable. Evidence that the measurement instruments possess these desired properties is the criterion to be met prior to their use in the field test. Again, the same audience would seem to be appropriate here as was appropriate prior to the use of the instruments in the pilot test.

As a result of use in the field test, data collection and measurement procedures should be revised/established to enable the appropriate use of the measurement instruments by qualified individuals in the educational research community. The audience most relevant here would be comprised of the consumer representatives.

The fourth issue noted is concerned with determining whether or not different effects (e.g., variations in students' acquisition of the objectives) result from any alternative procedural options or variations on a given procedure (i.e., degree of implementation effects). The criterion to be met prior to hot house would appear to demand some methodology for identifying deviations from use procedures as specified by the developer (i.e., use procedures outlined in the teacher's guide accompanying a particular package or unit). The most relevant audience for determining fulfillment of this criterion would be comprised of the developer, evaluation advisors, and substantive advisors.

Prior to use of the product in the pilot test, the procedural deviations noted in hot house should be identified and classified in order to facilitate the use of some systematic means of recording and appraising the degree of classroom implementation, which would enable a more precise specification of procedural deviations of interest. At this stage the audience would be

comprised of the developer, the substantive and evaluation advisors, and the institution, which would presumably be concerned with the kinds of procedural deviations noted and the nature of support service requirements involved in appraising degree of implementation effects.

Prior to entering the field test, the criterion to be met requires the establishment of some methodology for relating any degree of implementation effects (noted during pilot) to dependent variables as specified in terms of the stated objectives (see III A). The evaluation advisors and the developers would seem to comprise the most relevant audience for evaluating evidence concerning this criterion due to: 1) the methodological considerations involved, and 2) the extent to which the developer's objectives are affected by variations in degree of implementation procedures.

The prerequisite criterion for public diffusion would require that any information concerning differential implementation effects noted in the field test be made available to prospective users. The most relevant audiences would be the consumer representatives (who obviously would want to know what effects implementation variations might have on the populations they represent), and the sponsor (who seeks justification for committing the resources necessary for product development and testing).

The next major issue we have identified is concerned with whether or not there are important differences in the accomplishments of individual students that warrant the investigation of output to prerequisite variable relationships. Information regarding these types of relationships obviously cannot be obtained systematically prior to the initial use of the product during hot house. Prior to entering hot house, however, the

criteria to be met for acquiring the relevant information demand some specification of the prerequisite variables of interest (e.g., age, sex, SES, I.Q., measures of relevant skills, etc.) and evidence of the integrity of any measures used to assess the prerequisite variables of interest. The audience attending to these criteria would be comprised of the substantive advisors (who should advise on the possibility of such relationships), the developer (who should have to specify the nature of any such relationships to the consumer), and the evaluation advisors (on methodological grounds).

Prior to the pilot test, the criteria would require: 1) the selection (based on evidence acquired in hot house) of the prerequisite variables to be used in future studies and, 2) a suitable experimental design which would appropriately handle the prerequisite variables selected for study. The composition of the audience would be the same as for the criteria prior to hot house for the same reasons.

As a result of the application of a suitable assessment strategy in the context of an appropriate experimental design, any evidence of individual differences in the outcome variable set which are contingent on prerequisite status should become apparent. If such evidence exists, the criterion to be met prior to the field test would require the establishment of an experimental design which relates prerequisite and dependent variables to enable a more systematic test of these relationships. Again, the audience composition is the same.

As a result of the information gathered during the field test, some recommendations concerning appropriate use patterns (to capitalize on desirable prerequisite - dependent variable relationships) should be made available to potential consumers. This s the criterion we have identified that should be met prior to public diffusion. The audience would be comprised of the consumer representative (who would evaluate the adequacy of the information presented according to the information requirements of those he represents) and the sponsor (who, again, is presumably interested in this aspect as it relates to specifications of the nature of the product he has funded).

The next issue to be considered is the determination of whether there are important longitudinal effects that may result from use of the product beyond any immediate effects (attainment of objectives) expected to occur as a result of specified use procedures. If this issue is important, then the criterion to be met by the developer prior to any initiation of product development would require the identification of any hypothesized longitudinal effects. The audience attending to this criterion would be comprised of the substantive advisors (longitudinal effects are substantive issues) and the sponsor (the investigation of such effects may require substantial increases in funding).

No further criteria would seem to be applicable for adequately contending with this issue prior to the field test since, prior to this point in time, the product has typically undergone some revisions. Should such revisions be substantial (e.g., revisions concerning objectives, use procedures, content), there could be no precise specification of the independent variable (i.e., the product) and, consequently, any efforts

to systematically investigate hypothesized longitudinal effects would be fruitless.

The most opportune time to begin the investigation of these effects (concurrent with program development as opposed to a totally external summative effort) is during the field test when the most advanced version of the product is available. The criteria to be met for warranting such investigative efforts would obviously require the establishment of an experimental design for investigating the hypothesized effects in the field test population. Data collection procedures and measurement instruments also must be available for the observation of postulated effects. The most appropriate audience at this point might consist of the developer (who is obviously interested in how his postulated effects are to be (identified), the substantive advisors (to review the correspondence between the nature of the revised product and postulated effects), and the evaluation advisors (to review instrumentation and design components). In most instances, evidence of longitudinal effects gathered during the field test would be indicative rather than conclusive, since the time alloted for field tests will probably not be sufficient to allow complete investigation of postulated effects prior to public diffusion of the product. One can gather, however, the available information, synthesize it, and make some recommendations for longitudinal studies to the educational research community.[4] This is

---

[4]Additional efforts to investigate the field test sample to acquire additional evidence after public diffusion might be undertaken jointly by the developer and institution. We feel as Karl (1970) does, however, that such ad hoc efforts are typically not undertaken due to: 1) the lack of additional funds; 2) the developer and project personnel are off to contend with newer, more attractive projects looming on the horizon. We have chosen to label this latter phenomenon the "Horizon Syndrome."

perceived as the criterion to be met prior to public diffusion. The audience would be comprised of the consumer representative (for obvious reasons), the sponsor (since this information relates to specifying the nature of the product), the substantive advisors (as this information relates to substantive issues), and the evaluation advisors (who review the credibility of the information to be presented in terms of the analytic procedures used).

The last of the issues identified in the product worth dimension deals with the determination of any positive or negative unanticipated consequences (especially cross-curricular) occuring as a result of product use. Any systematic investigation of such consequences would require the inclusion of some means (e.g., questionnaires, participant observation, interviews, unobtrusive measures, etc.) for their discovery in the overall assessment strategy implemented in the hot house trials. The developer and evaluation advisors are perceived as the most relevant audience due to: 1) the fact that the occurance of unanticipated consequences might substantially influence product revisions, and 2) the methodological considerations involved.

Prior to entering the pilot test, any unanticipated consequences identified during hot house related to the product's use should be built into future research designs for more controlled investigation. Obviously some initial product revisions would be undertaken to overcome any negative consequences tentatively identified during hot house, but procedures should be developed to continue attempts to identify unanticipated consequences

during pilot and field trials. The audience attending to these criteria should be comprised of the developer and substantive advisors (product use effects and subsequent revisions are of major concern), the evaluation advisors (methodological considerations), and the institution (additional support services may be required and then must share responsibility for any negative consequences).

Upon termination of the pilot test, necessary product revisions should be undertaken to eliminate any negative, and maximize any positive, unanticipated consequences initially identified during hot house and more systematically investigated during pilot. Additional unanticipated consequences noted during the pilot test should be built into the field test research design for more precise investigation. The audience attending to these criteria is the same audience attending to the pilot test criteria and for substantially the same reasons.

Prior to public diffusion, the final version of the product should be revised to take into account the identified consequences. Both positive and negative consequences discovered in earlier trials, as well as the efforts undertaken to deal with them, should be reported to the educational research community. Since unanticipated consequences can demand revisions which critically effect the nature of the product (i.e., it's content, materials, activities, objectives, measures, etc.) and specify limitations on it's use patterns, all audiences (especially the consumer representatives) should evaluate the adequacy of the information presented in fulfillment of these criteria.

UTILITY OF THE PROCEDURE

The specification of criteria in the manner we have proposed is certainly an involved and time-consuming task. Is it worth the effort? Do the anticipated beneficial consequences warrant the expenditures? In order to experimentally investigate the question, one would need to compare the results of two attempts to generate similar products, one of which used this approach and the other of which did not, while controlling other variables such as the quality of the personnel, the facilities, etc. Clearly such a study is not worth its cost.

On pragmatic grounds, however, we believe that a case can be made for the use of this or a similar procedure. Perhaps the most frequent complaint of curriculum development people is that in order to continue their work they are subject to frequent reviews, site visitations, annual reports and the like. These events are quite disruptive in the sense that the principal mission of the program must be abandoned momentarily while preparations are made to meet the needs of the sponsoring agency. As someone has put it, in order to see if the rose is growing you pull it from the soil and examine it's roots, at least once a week. We doubt that many seriously would argue that the sponsor is not entitled to investigate the progress of funded projects. What is needed is a mechanism for insuring that the investigation attends to the right issues and does not serve merely to generate anxiety and inane show and tell sessions.

If we can view the developmental/evaluation stages as milestones in the life of a product, then it is possible to specify a procedure for reviews that might be minimally disruptive. Given the product specifications

of the developer, the expected milestones within a calendar period (e.g., three hot house and four pilot trials in a calendar year), and the criteria for advancement for the given product stages, it should be possible for the sponsor to examine merely the discrepancies between anticipated and actual milestone accomplishment in order to determine whether or not the developer's efforts are worthy of continuing support.

Moreover, one could use the criteria postulated to determine whether the "fit" between the plans of the developer and the perceived needs of the sponsor was sufficient to warrant funding. The general goals might be similar, but if, for example, the developer would weight very highly criteria related to student affective consequences and he plans to develop his product and then engage in the process of determining how teachers need to be educated in its use, whereas the sponsor perceives the need, in this discipline, for a product that is low cost, can be used by teachers without special training, and emphasizes cognitive outcomes, then a lack of fit exists and negotiation must lead to a resolution or to the rejection of the proposal. The advantage is that once criteria are specified and agreed to, the efforts of the developer must be judged on those terms, not on some arbitrary fluctuating grounds.

In terms of the operation of a project, the specification of the criteria should promote more appropriate research and evaluation. There is no justification for collecting information irrelevant to the decision-making process. Unless one is aware of the criteria for decision making, however, the collection of the relevant information is at best haphazard. Evaluating the evaluator is possible only if one knows what the evaluator's responsibilities really are.

The complete specification of all criteria for every stage may be virtually impossible at the beginning of a project. We think, however, that it is possible to identify in some detail the criteria for at least the current and the next in-sequence stage and to identify alternative general criteria for later stages. As the project proceeds it should be possible to flesh out the charts so as to inform others of one's intentions well in advance of a given trial.

In the final analysis, one might state that the purpose of any such procedure is improved communication and its utility rests in the necessity for this communication. We perceive it to be absolutely necessary if evaluation at all levels is to become more precise and valuable.

SUMMATIVE EVALUATION - CONSUMER PROTECTION

In the tables we have presented, evaluation ceases when the product enters the diffusion stage. It seems to us that it is at this point that formative evaluation is ended. The product is presumably a finished entity, and while revision in the form of a second edition is possible or even probable, the product as it stands is available to the purchasing public. It is undoubtedly considered by those responsible for its development as something distinct and meritorious, and is promoted as such by them. Indeed, they may have firm grounds for so doing, assuming that the evaluation up to this point has been well conducted.

Yet, regardless of how thoroughly accomplished it might be, evaluation done in-house, while the product is still in development, is nonetheless formative. The distinction between formative and summative is perhaps not entirely discrete. Elements of what might be considered summative

evaluation can be located in the latter stages of our strategy. The distinction nonetheless is critical. At some point evaluation must be externally conducted. The dangers of cooptation and vested interest on the part of an in-house evaluator are very real, even though scrupulous attention is paid to the issue of evaluator independence. Moreover, evaluation for the consumer must be goal-free in the sense that the multiple consequences of product use and not merely those objectives identified by the developer are studiously examined.

It seems to us that the salient characteristic of true summative evaluation then might be conceived of as consumer protection. Drug companies, for example, tout the virtues of their patent medicines on the basis of their own research, some of which is unquestionably sound. Yet the intelligent consumer relies on other sources for his information. The reports of the Food and Drug Administration inform him that the product has met certain standards (not stringent enough, perhaps, but reasonably well defined). Reports of the Consumer's Union and other similar groups provide him with comparison data on certain frequently used classes of drugs. For more precise information, he may rely on a physician who has access to specialized data sources and reports and is presumed competent to interpret them.

At the moment no such formal mechanisms for consumer protection in the educational products field exist. The cost of initiating this service would be quite high; yet, we feel that it is already overdue and worth the expenditure. We can see a three-tiered process operating wherein: 1) the product developer establishes the criteria in conjunction with the funding

agency and the other interested parties and formatively evaluates his product in those terms; 2) some agency or center established by the federal government examines the product to insure that: a) it does indeed stand up to the claims made for it (the truth in advertising issue); b) there are no adversive consequences from use (for example, the thalidomide or withdrawal effects discussed by Weiss [1971]); and c) that a conscientious, external, goal-free evaluation is conducted and reported so that the intelligent consumer can make informed adoption decisions; and 3) finally local education agencies (LEA) should be in a position to conduct reasonably competent research on products in the field in order to determine whether or not these products perform the specific function desired. The best mechanism for this might well be an institute established by a consortium of L.E.A.'s which can provide the necessary expertise and technical capability to undertake such research. The authors think that the Institute for Educational Research in Downers Grove, Illinois is a worthy example.

BIBLIOGRAPHY

Astin, A. W. and Panos, R. J. "The Evaluation of Educational Programs." Educational Measurement. 2nd ed. Edited by Robert L. Thorndike. American Council on Education, Washington, D.C., 1971.

Baker, Robert L. "Curriculum Evaluation." Review of Educational Research, XXXIX, 3 (June, 1969), p. 339-358.

Berlak, Harold. "Values, Goals, Public Policy and Educational Evaluation." Review of Educational Research, XL, 2, (April, 1970), p. 261-278.

Campbell, Donald T. and Stanley, Julian C. "Experimental and Quasi-Experimental Designs for Research on Teaching." Handbook of Research on Teaching. Edited by N. L. Gage. A project of the American Educational Research Association. Chicago: Rand McNally, 1963.

Cronbach, L. J. "Course Improvement Through Evaluation." Teachers College Record. 64, 1963, p. 672-683.

Eash, M. "Developing an Instrument for the Assessment of Instructional Materials (Form IV)." Paper presented at the AERA Annual Conference, Division B, Minneapolis, Minnesota, March, 1970.

Grobman, Hulda. "Evaluation Activities of Curriculum Projects." AERA Monograph Series on Curriculum Evaluation. Chicago: Rand McNally, 1968.

Guba, E. G. and Stufflebeam, D. L. "Evaluation: The Process of Stimulating, Aiding, and Abetting Insightful Action." An address delivered at the Second National Symposium for Professors of Educational Research, Boulder, Colorado, November 21, 1968.

Karl, Marion. "An Example of Process Evaluation." Paper presented at the AERA Annual Conference, Division B, Minneapolis, Minnesota, March, 1970.

McDonald, Barry. Evaluation of the Humanities Curriculum Project: A Wholistic Approach. Centre for Applied Research in Education, University of East Anglia, England, December, 1970.

Provus, Malcolm. Discrepancy Evaluation for Educational Program Improvement and Assessment. Berkeley, California: McCutchan Publishing Co., 1971.

Russell, Howard, ed. Evaluation of Computer Assisted Instruction Program. St. Louis, CEMREL, Inc., 1969.

Scriven, Michael. "The Methodology for Evaluation." Perspectives of Curriculum Evaluation: AERA Monograph Series on Curriculum Evaluation. Chicago: Rand McNally, 1967, p. 39-83.

Scriven, Michael. "Who Evaluates the Evaluators." (Mimeo, 1971).

Stake, Robert E. "Objectives, Priorities and other Judgement Data." Review of Educational Research, XL, 2, (April, 1970), p. 239-260.

Stufflebeam, Daniel L. "The Use of Experimental Design in Educational Evaluation." Journal of Educational Measurement, VIII, No. 4 (Winter, 1971), p. 267-274.

Tyler, Louise and Klein, M. Frances. Recommendations for Curriculum and Instructional Materials. University of California, Los Angeles, 1967.

Tyler, Ralph W., ed. Educational Evaluation: New Roles, New Means. The Sixty Eighth Yearbook of the National Society for the Study of Education, Part II. Chicago, 1969.

Weiss, Joel. "Formative Curriculum Evaluation: In Need of Methodology." Paper presented at the Annual Conference of the AERA, New York City, February, 1971.

Westbury, Ian. "Curriculum Evaluation." Review of Educational Research, XL, 2, (April, 1970), p. 239-260.

Wittrock, M. C. "The Experiment in Research on Evaluation of Instruction." Center for the Study of Evaluation of Instructional Programs, Working Paper No. 2, December, 1966. University of California, Los Angeles.

Wright, William J. "Evaluation in the Aesthetic Education Program: Critique and Proposal" (Mimeo, 1972).